Lightweight Indoor Positioning System Based on Multiple Self-Learning Features and Key Frame Classification

Chenzhe Wang [1], Kai Bi [1], Bianli Zhao [1], Ming Li [1], Yujia Chen [2], Shiliang Tao [3], Juntao Yang [4]

[1] National Geomatics Center of China, Beijing, China - wangchenzhe@ngcc.cn, 24677958@qq.com, zbl@ngcc.cn, liming@ngcc.cn
[2] North China University of Science and Technology, Tangshan, China - chenyujia@ncst.edu.cn
[3] Baidu.Inc, Beijing, China - 505165517@qq.com
[4] Shandong University of Science and Technology, Qingdao, China - jtyang@sdust.edu.cn

Keywords: Indoor Positioning System, Key Frame Classification, Convolutional Neural Network, Feature Point Recognition, SuperPoint, MobileNet V3-Small.

Abstract

Traditional indoor positioning technologies mostly require advanced installation of hardware devices, resulting in high costs and long-term maintenance. With advancements in image recognition and deep learning technologies, indoor visual positioning based on image recognition has become increasingly mature. This method offers the benefits of low cost and does not require additional hardware installation. However, it still has inherent defects, such as cumbersome data collection, complex algorithms, and universality. To minimize indoor information pre-collection cost, improve versatility, and enable rapid deployment in low-performance mobile devices, this paper proposes a lightweight indoor positioning system based on multiple self-learning features and key frame classification. The system is divided into two stages: preprocessing and real-time positioning. In the preprocessing stage, image information is collected for the entire indoor environment, and a key-frame recognizer is trained based on the image information. Simultaneously, an environmental feature information database is established. In the real-time positioning stage, the system first uses mobile devices such as smartphones to obtain real-time video streams. A key frame recognizer based on convolutional neural networks identifies key frames in each video stream frame, thereby obtaining approximate positions for rough positioning. Second, feature points are identified in each frame of the video stream and matched with feature points with location information in the built environmental feature information database to calculate precise positions for fine positioning. It has significant optimizations compared with conventional visual solutions in terms of preprocessing data collection, algorithm performance consumption, and versatility.

1. Introduction

Recently, location-based services have received widespread attention and promoted the rapid development of positioning technology. The inability of the global navigation satellite system (GNSS) technology to solve the positioning problem in indoor scenarios has led to the emergence of indoor positioning technology, with potential applications influencing people's daily lives, including exhibitions, airports, train stations, and large shopping malls. The currently main indoor positioning technologies mainly include Wi-Fi, Bluetooth beacons (Yang et al., 2022), pedestrian dead reckoning (PDR), geomagnetism, and ultra-wideband (Yao et al., 2022). Wi-Fi primarily uses fingerprint comparison algorithms, which require equipment deployment in advance and collecting a large amount of fingerprint information (Gholami et al., 2019). The latest solution is to adopt crowdsourcing of signal sources and add location semantics to reduce the preprocessing of signal fingerprint collection; however, it is still limited by uneven crowdsourcing locations; therefore, it cannot solve the precision problem of all locations. Bluetooth uses fingerprint matching or similar base station signal trajectory matching methods with higher precision but still requires equipment deployment. Although the PDR algorithm does not require advanced equipment deployment, a cumulative error occurs that cannot be eliminated, and it must be combined with Bluetooth or Wi-Fi for absolute position correction (Huang et al., 2022).

Visual-based indoor positioning technologies have matured with image recognition and deep learning development. It is inexpensive and does not require additional equipment, making it more advantageous than traditional solutions. However, there are also specific defects in current visual-based indoor positioning, such as the indoor positioning algorithm based on simultaneous localization and mapping (SLAM) and fusing visual odometry, which have high complexity and low efficiency and cannot meet real-time requirements (Mansour et al., 2023). Schemes based on feature point recognition and position conversion through homography require large amount of image data in advance; their universality is not high, and positioning continuity cannot be guaranteed (Hung et al., 2019). Schemes based on deep learning object recognition have no additional hardware dependencies but cannot accurately estimate positions (Ciou and Lu, 2019). Therefore, to address the limitations of existing indoor positioning technologies, this study proposes a lightweight indoor positioning technology based on multiple self-learning features and key frame classifications. This technology only requires video stream images captured by mobile terminals to obtain real-time and accurate indoor positions and optimizes performance consumption and universality efficiently (Basri and Elkhadimi, et al., 2020).

2. System Process Overview and Core Steps

The implementation process of the indoor positioning system based on the vision mentioned in this paper is shown in Figure 1, which is divided into a data preprocessing stage (blue part on the right) and a real-time positioning stage (green part on the left). In the data preprocessing stage, key frame positions within the indoor environment are carefully selected, and an adequate number of images are captured at these positions. After processing, classifier models and feature point descriptions for each key frame are obtained. During the real-time positioning stage, the key frame classifier is used to identify the matching key frame in the current frame for rough localization. Precise

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-373-2024 | © Author(s) 2024. CC BY 4.0 License.

current position is then determined through feature point matching and triangulation principles for fine localization.

The detailed steps are as follows (Section 2.1 and Section 2.2).

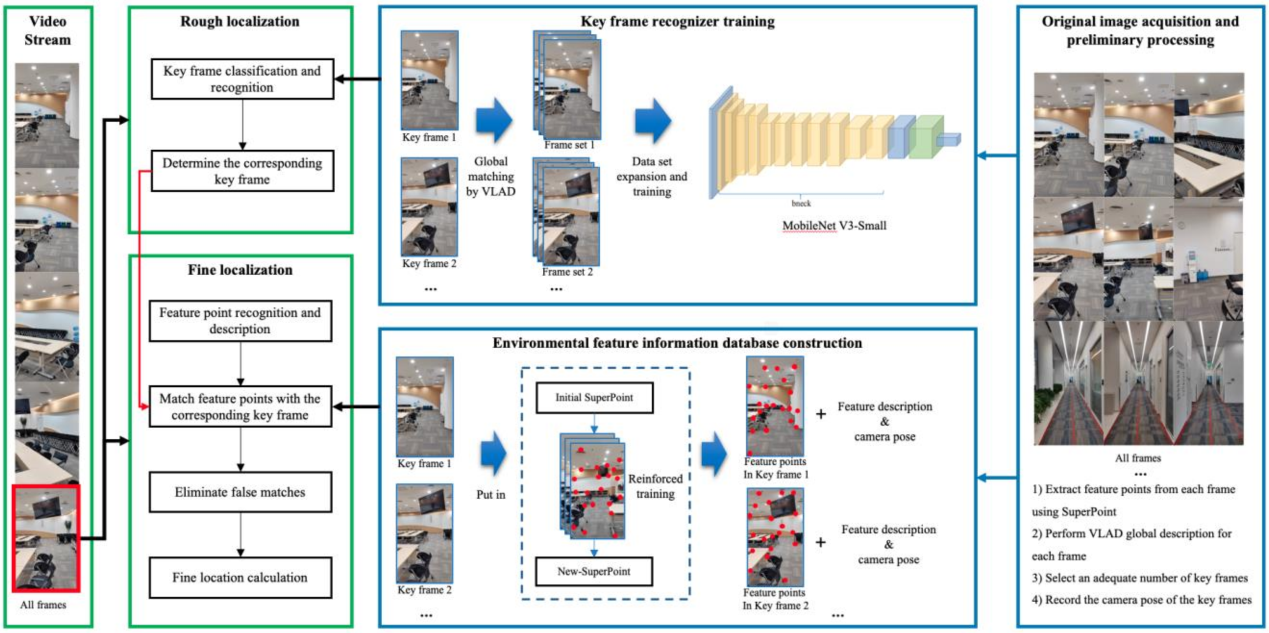
Video Stream
All frames
Rough localization
Key frame classification and recognition
Determine the corresponding key frame
Fine localization
Feature point recognition and description
Match feature points with the corresponding key frame
Eliminate false matches
Fine location calculation
Key frame recognizer training
Key frame 1
Key frame 2
...
Global matching by VLAD
Frame set 1
Frame set 2
...
Data set expansion and training
bneck
MobileNet V3-Small
Environmental feature information database construction
Key frame 1
Key frame 2
...
Put in
Initial SuperPoint
Reinforced training
New-SuperPoint
Feature points In Key frame 1
Feature points In Key frame 2
...
+
Feature description & camera pose
+
Feature description & camera pose
Original image acquisition and preliminary processing
All frames
...
1) Extract feature points from each frame using SuperPoint
2) Perform VLAD global description for each frame
3) Select an adequate number of key frames
4) Record the camera pose of the key frames

Figure 1. Workflow chart of indoor positioning system based on deep learning and feature point recognition.

2.1 Data Preprocessing Stage:

2.1.1 Original Image Acquisition and Preliminary Processing: The indoor scene was comprehensively imaged using video equipment. However, too many images were not required. Only 3-4 repetitions at the same location were required. After the acquisition, the video stream was divided into frames, and the SuperPoint (a self-supervised method for interest point detection and description, published in 2018) feature points and vector of locally aggregated descriptors (VLAD) global feature descriptors were extracted from each frame. Finally, N key frames with distinct and evenly distributed features were selected, and the camera poses information was recorded during key frame capture (Jégou et al., 2010).

2.1.2 Key Frame Recognizer Training: For each key frame, global matching was performed on the original image set to obtain a similar image set that belonged to each key frame. All image sets were subjected to augmented transformations (such as grayscale, filtering, rotation, and affine transformation) to obtain the expanded training set. The training set was used for transfer learning on MobileNet V3-Small to obtain a key-frame recognizer.

2.1.3 Environmental Feature Information Database Construction: The feature points from all frames in the original image acquisition were imported into the initial SuperPoint model for enhanced training and fine-tuning, resulting in the creation of a new model named New-SuperPoint This model represents a feature point recognition system achieved through targeted multi-round self-learning tailored specifically for indoor scenes, marking a notable innovation within this paper. New-SuperPoint proves to be more apt for the current indoor environment compared to directly using SuperPoint. New-SuperPoint was used to identify and describe all keyframes, and the camera pose of each keyframe was added to form an environmental feature information database.

2.2 Real-Time Positioning Stage:

2.2.1 Rough Localization: Based on the real-time video stream, each image frame was sent to the trained key-frame recognizer to obtain the approximate position of the frame and its corresponding key-frame image.

2.2.2 Fine Localization: Each video stream frame was processed using New-SuperPoint for feature point recognition and description and matched the corresponding key frame in the environmental feature information database. Accurate positional information was obtained through affine transformation and homographic decomposition.

3. Detailed Explanation of Key Technologies

3.1 Key Frame Recognizer Construction

3.1.1 Feature Point Recognition Based on Convolutional Neural Networks: Traditional feature point extraction was performed manually using fixed algorithms (such as SIFT, etc.). Due to the difficulty in fixed-algorithm optimization, most have poor general adaptability and are only suitable for partial scenes. Moreover, if more accurate feature points are desired, the algorithm's complexity becomes extremely high, making it challenging to satisfy the performance requirements of mobile terminals. To solve these problems, this study adopted SuperPoint, a feature-point recognition algorithm based on convolutional neural networks. SuperPoint is a feature point detection and description method based on self-supervised training. It consists of an encoder and two decoders and can simultaneously produce feature points and descriptors (DeTone et al., 2018), as shown in Figure 2.

1) The feature encoder: This encoder utilizes a structure similar to a VGG (the network model proposed by the Visual Geometry Group at the University of Oxford in 2014) for feature extraction and dimensionality reduction. It comprises convolutional layers, spatial downsampling through pooling, and nonlinear activation functions. Conv2d in Figure 2 stands for 2D convolution. The encoder uses three max-pooling layers, which results in a two-fold downsampling in the spatial dimensions three times. Assuming that the input image size is $H \times W$, the tensor map after the encoder has dimensions of $H_C \times W_C \times F$, where $H_C = H/8$, $W_C = W/8$, and F is the number of channels. This indicates that each pixel point on the tensor map represents a local image block of 8×8 pixels in the original image.

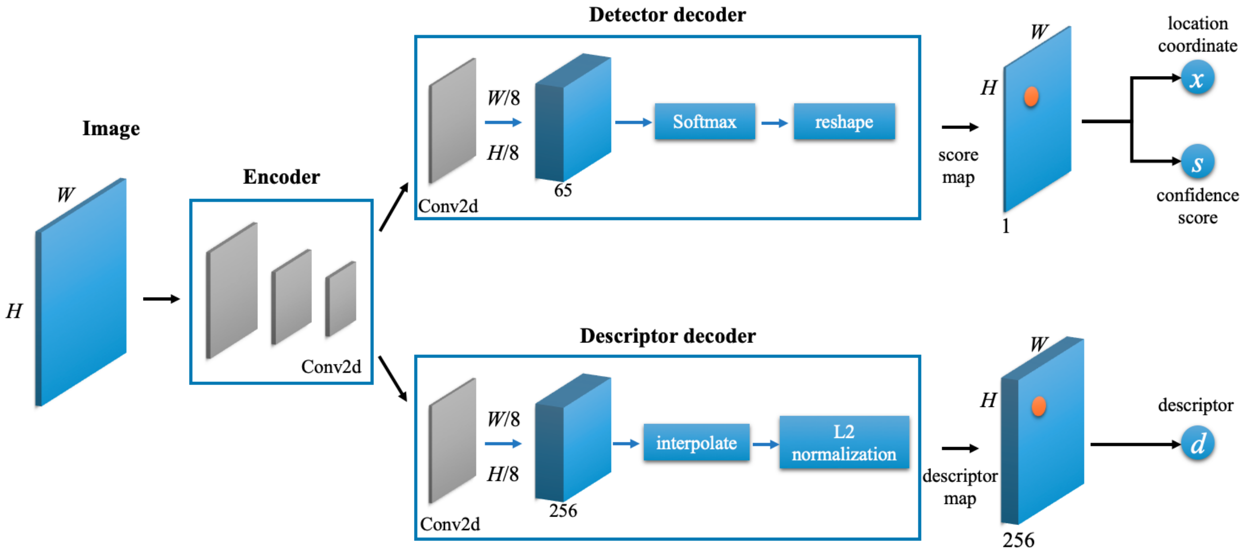

Figure 2. Feature point extraction and description process using SuperPoint.

2) Feature-detection decoder: The tensor map output from the encoder passes through two layers of convolutional modules in the decoder. The intermediate tensor has a dimension of $H_C \times W_C \times 65$, where the first 64 channels are used to predict the probability of each pixel in the 8×8 local region of the original image being a feature point, and the last channel represents the probability of there being no feature points in this region. The intermediate tensor undergoes a Softmax operation, removing the last channel, and the reshape operation is used to change the tensor's dimension from $H_C \times W_C \times 64$ to $H \times W$, obtaining a confidence score map of feature points of the original image size. According to the confidence scores s, pixels with scores greater than the threshold s_{th} are designated as feature points, and the corresponding pixel positions are set as the coordinates x of the feature points.

3) Feature description decoder: This decoder consists of two basic convolutional layers, and the output tensor has dimensions $H_C \times W_C \times 256$. By directly utilizing bilinear interpolation, the dimensions were changed to $H \times W \times 256$. Finally, L2 normalization was applied to the channels of each pixel point to obtain the normalized descriptors. Sampling-based on feature point locations yields features point descriptors d.

3.1.2 Global Feature Description and Matching: It was necessary to retrieve frame sets with high similarity to each key frame from the original image set as the training set to train the key-frame recognizer. Similar frame retrieval based on simple feature points is insufficient, and each frame's features must be clustered into a single global descriptor vector. This study used the VLAD algorithm for global descriptor generation and similar-frame retrieval.

The VLAD algorithm assumes that each frame image has $N \times D$-dimensional local feature points (where N may be relatively large and the number of features in each image varies). It is necessary to obtain a $K \times D$-dimensional feature (where K is a specified number, such as 128 dimensions) representing the global image from the $N \times D$-dimensional features. The main process is as follows.

1) K-means clustering is performed on all $N \times D$-dimensional local features to obtain K cluster centers, denoted as C_k.

2) The $N \times D$-dimensional local features are encoded into a global feature $\boldsymbol{V}$ with a feature vector dimension of $K \times D$, where $k \in$ K and $j \in$ D. The formula used is as follows:

$$\boldsymbol{V}(j,k) = \sum_{i=1}^{N} a_k\,(x_i)(x_i(j) - c_k(j)) \tag{1}$$

Where x_i is the i-th local image feature
c_k is the k-th cluster center
x_i and c_k are D-dimensional vectors
$a_k(x_i)$ is a binary function that equals 1 if and only if x_i belongs to the cluster center c_k, and 0 otherwise

Finally, the global features of the image after dimension reduction can be obtained, and all images similar to the key frame can be retrieved by setting a suitable Euclidean distance threshold. All retrieved frame sets will be utilized as a training dataset for the training of the key frame recognizer.

3.1.3 Key Frame Recognition Model: The main system in this study was deployed on mobile devices, and the time required by the algorithm was high; therefore, MobileNet V3-Small was selected as the basic network structure. MobileNet V3, with network structure optimization, achieves higher accuracy than most large neural networks, with fewer parameters and lower computational cost (Shi et al., 2020). The latest version of MobileNet, V3-Small, has a computational speed of 22 ms, significantly faster than most large neural networks. MobileNet V3-Small has 12 unique Bneck layers, one standard convolution layer, and two pointwise convolution layers. It has the following characteristics (Howard et al., 2019).

1) With the depth-wise separable convolution of MobileNetV1, the number of parameters and computations is lower than that of the standard convolution while maintaining a similar accuracy.

2) With the linear bottleneck inverse residual structure of MobileNetV2, this structure can reduce the number of parameters and convolution calculations compared to the standard convolution, optimizing the network in both spatial and temporal dimensions.

3) By adding lightweight attention models (squeeze-and-excitation modules), the network can assign larger effective weights to the input features and smaller weights to ineffective or less effective features.

4) Use the h-swish activation function. Using this activation function in experiments with Google AI improved the efficiency by approximately 15%.

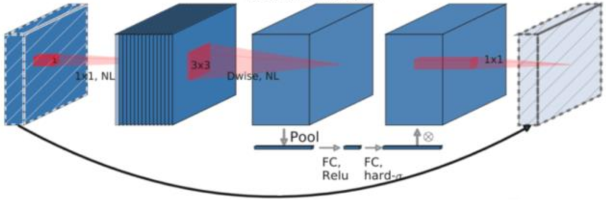

Figure 3. The unique bneck structure of MobileNet V3.

3.1.4 Key Frame Recognition Transfer Learning: In this study, the pretrained model parameters were obtained by transferring the well-trained weights of MobileNet V3 to the ImageNet dataset. The pretrained model parameters were fine-tuned by training specific layers and freezing other layers to obtain the desired keyframe recognizer.

3.2 Environmental Feature Information Database Construction

3.2.1 Multiple Self-Learning Reinforced Training for New-SuperPoint: Although the original SuperPoint recognition model has significantly improved efficiency compared with traditional algorithms, it still has poor detection results in indoor areas that have not been learned because of its deep learning approach. To solve this problem, this study will continue to use all the feature points obtained from the original SuperPoint recognition of indoor area images as the training set input to the model for parameter fine-tuning and obtaining a more targeted new SuperPoint. As SuperPoint is a self-learning model, the new SuperPoint used in this study is a multiple self-learning model.

3.2.2 Environmental Feature Information Database Construction: Using New-SuperPoint to identify feature points in all key-frame images, we obtained the corresponding feature points and descriptive information. By combining the positions of the key-frame images and the pose information of the camera at the time of capturing each key-frame, we obtained the final environmental feature information database.

3.3 Indoor Coordinate Calculation

In the rough localization stage, a key-frame recognizer is used to determine the approximate position of each frame in the video stream and to obtain the corresponding key-frame images. In the fine localization stage, to calculate precise coordinates, it is necessary to recognize and describe the feature points for each frame in the video stream. After obtaining the feature points of each frame in the video stream, they were matched with the feature points in the environmental feature information database of the corresponding key frame, and the precise position was calculated accordingly.

3.3.1 Feature Point Matching: In this study, the SuperGlue method was used for feature-point matching. Published in 2020, SuperGlue is a deep learning network framework designed for graph matching of large-scale feature points. It utilizes the attention mechanism of Transformer to adaptively enhance the global information of feature points, thereby improving both the number and accuracy of matches. The self-crossing mechanism unique to SuperGlue is superior to traditional feature-matching algorithms in terms of efficiency and accuracy.

3.3.2 Accurate Position Calculation: Stable point pairs were selected from the matched point pairs, and the homography matrix H (Equation (2)) was calculated. The rotation and translation matrix are obtained through SVD decomposition, and the camera attitude angle $\theta_{attitude}$ is obtained. The stable feature point sets were randomly jumped, and the differences were calculated. The reverse weights are assigned based on the size of the differences, and the weighted average of the feature point position distance difference between the template image and the video stream image $\delta_{distance}$ is obtained. The known camera focal length f, camera attitude angle $\theta_{attitude}$, template image position coordinates P_0, and template image shooting distance z_0 are used to calculate the distance between the mobile phone and the current template image, and the precise position P (Equation (3)) is obtained.

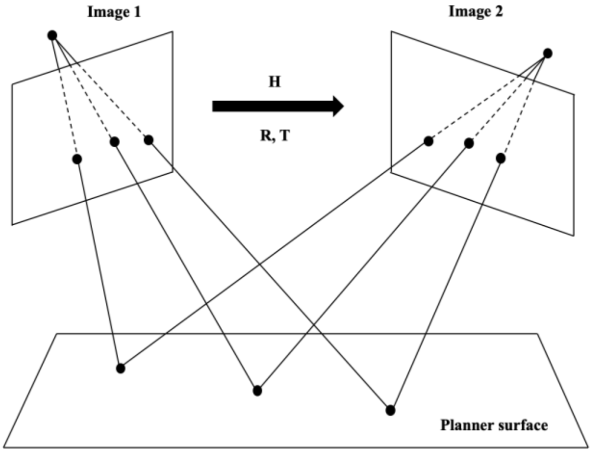

Figure 4. Camera pose estimation based on homography matrix computation.

$$\boldsymbol{H} = \boldsymbol{K}(\boldsymbol{R} + \boldsymbol{T}\frac{1}{d}\boldsymbol{N}^T)\boldsymbol{K}^{-1} \tag{2}$$

where $\boldsymbol{K}$ represents the internal parameters of the camera
d is the center distance
$\boldsymbol{N}$ is the normal vector of the camera plane
$\boldsymbol{R}$ is the external rotation matrix of the camera
$\boldsymbol{T}$ is the external translation vector of the camera.

$$\boldsymbol{P} = \boldsymbol{P_0} + [(\delta_{distance}\, f + z_0)sin\theta_{attitude} \quad (\delta_{distance}\, f + z_0)cos\theta_{attitude}]^T \tag{3}$$

4. Experiment and Results

4.1 Experimental Environment and Hardware Platform

Experiments were conducted to validate the proposed indoor positioning technology and evaluate its feasibility and precision. Details of the hardware and software platforms used in the experiment, along with related parameters, are provided in Table 1.

Platform	Model	CPU	GPU
Training Platform	Samsung Galaxy Z Flip5	Qualcomm Snapdragon 8 Gen 2	Qualcomm Adreno 740
Experimental Platform	MacBook Pro 2020	Intel Core i5 1.4 GHz	Intel Iris Plus Graphics 645

Table 1. Parameters of the software and hardware used in the experiment.

In this study, an Android phone was used as the positioning terminal, and all models were trained on the PC side and then migrated to the mobile side. Using TensorFlow, we obtained the jar and pb files required for deployment to Android devices, and these files were placed in an Android project and configured with Gradle to deploy all models on the mobile Android side.

Furthermore, this study did not conduct experiments using a wide variety of smartphone models, as the performance and image capture capabilities of most smartphones are similar. The test smartphone used in this study has average performance, making it representative of the majority of smartphones available on the market.

4.2 Data Acquisition and Preprocessing

1) Images were acquired throughout the experimental area and feature points were identified for each frame.

2) Key frames with distinct and evenly distributed feature points were selected from the captured images, and the camera pose information was recorded when capturing the key frames.

3) The key-frame recognizer and environmental feature information database were constructed according to the method introduced in Section 2. The training sets underwent a range of augmented transformations (grey scale, contrast, filtering, rotation, and affine transformation) to obtain more accurate models before training the model.

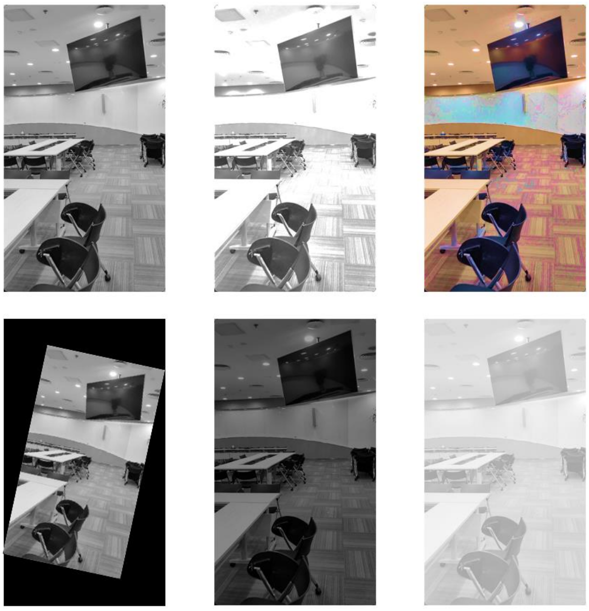

Figure 5. Template image training set obtained through enhancement transformation.

4.3 Experimental Result Analysis

Real-time positioning experiments were conducted at the experimental site with 50 trials on the same route to reduce the impact of accidental results. The average positioning trajectory of the experimental route is illustrated (Fig. 6). By comparison with the real trajectory; it can be observed that the positioning accuracy of the proposed method is relatively high, with more than 90% of the average error of the route controlled within 0.15 m. The average error of the positioning experiment is 0.12 meters, achieving sub-meter level precision. Compared to conventional indoor navigation methods, such as WIFI-based approaches, the precision advantage is obvious.

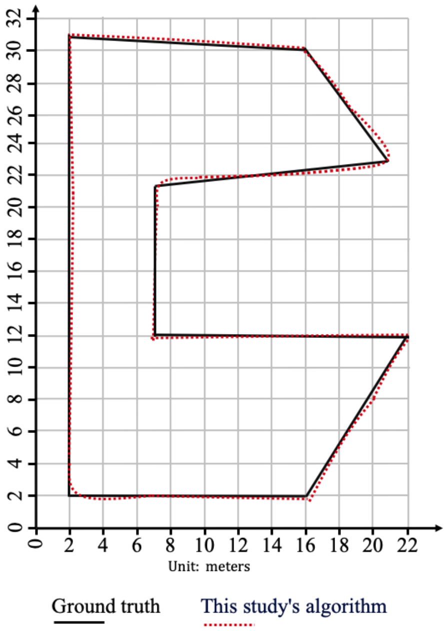

Figure 6. Comparison between ground truth and test track.

5. Conclusion

This study proposed an indoor positioning method based on multiple self-learning and key frame classification. This method utilizes a convolutional neural network-based key-frame recognizer and a feature point recognition model to achieve high-precision and real-time positioning capabilities. Experimental results showed that the positioning error of the proposed method was generally within 0.15 m and can process at least 12 fps. Compared with traditional indoor positioning schemes, this method eliminates the need for advanced hardware installation. It achieves higher positioning accuracy with lower performance consumption, making it suitable for deployment on mobile phones. Therefore, indoor positioning based on images has more potential for exploration.

In contrast to conventional vision-based indoor positioning, this study performs precise feature point matching after acquiring key frames. This approach results in more accurate positioning and is particularly well-suited for relatively similar indoor scenes.

However, if the indoor areas are extremely similar, the accuracy of the algorithm may decrease to some extent. In the future, we will continue to conduct in-depth research and optimize the algorithm to enhance its generalizability.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.42201486).

References

Yang, S., Liu, J., Gong, X., Huang, G., and Yin, F., 2022: An adaptive smartphone hybrid indoor positioning solution incorporating heterogeneous sensors, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVI-3/W1-2022, 243–248, https://doi.org/10.5194/isprs-archives-XLVI-3-W1-2022-243-2022.

Yao, H., Wang, X., Qi, H., and Liang, X., 2022: Tightly coupled indoor positioning using uwb/mmwave radar/imu, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVI-3/W1-2022, 323–329, https://doi.org/10.5194/isprs-archives-XLVI-3-W1-2022-323-2022.

Gholami, A., Pahlavani, P., Azimi, S., and Shakibi, S., 2019: Improved indoor positioning technique based on a geographic weighted regression, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLII-4/W18, 441–446, https://doi.org/10.5194/isprs-archives-XLII-4-W18-441-2019.

Huang, C.H., Chang, Y.F., Tang, Y.T., Tsai, M.L., and Chiang, K.W., 2022: An integrated indoor positioning algorithm for smartphone using pedestrian dead reckoning with magnetic fingerprint aided, I*nt. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLIII-B1-2022, 213–218, https://doi.org/10.5194/isprs-archives-XLIII-B1-2022-213-2022.

Hung, M. C., Liao, J. K., and Chiang, K. W., 2019: Indoor positioning based-on images aided by artificial neural networks, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLII-2/W13, 799–803, https://doi.org/10.5194/isprs-archives-XLII-2-W13-799-2019.

Mansour, A., Chen, W., Weng, D., Yang, Y., and Wang, J., 2023: Leveraging human mobility and pervasive smartphone measurements-based crowdsourcing for developing self-deployable and ubiquitous indoor positioning systems, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLVIII-1/W2-2023, 1119–1125, https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1119-2023.

Ciou, J.M. and Lu, E. H.C., 2019: Indoor positioning using convolution neural network to regress camera pose, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W13, 1289–1294, https://doi.org/10.5194/isprs-archives-XLII-2-W13-1289-2019.

Basri, C. and Elkhadimi, A., 2020: A review on indoor localization with internet of things, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLIV-4/W3-2020, 121–128, https://doi.org/10.5194/isprs-archives-XLIV-4-W3-2020-121-2020.

Jégou, H., Douze, M., Schmid, C., Pérez, P., 2010: Aggregating local descriptors into a compact image representation, *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*., 10.1109/CVPR.2010.5540039.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018: SuperPoint: Self-Supervised Interest Point Detection and Description, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*., 10.1109/CVPRW.2018.00060.

Shi, C., Xia, R., Wang, L., 2020: A Novel Multi-branch channel expansion network for garbage image classification, *IEEE Access*., 2020, PP(99): 1-1.

Howard, A., Sandler, M., Chu, G., 2019: Searching for mobilenetv3, *Proce edings of the IEEE/CVF International Conference on Computer Vision*., 2019: 1314-1324.